An Evaluation of Tools for Post-Editing Research: The Current Picture and Further Needs

Lucas Nunes Vieira
Newcastle University
School of Modern Languages, OLB
Newcastle upon Tyne, UK
l.nunes-vieira@ncl.ac.uk

Abstract

This paper presents a comparative evaluation of four tools that can be used to collect user activity data (UAD) in machine translation post-editing (PE) research: Tobii Studio, Translog-II, TransCenter, and PET. These tools are analysed here based on empirical data as a way of providing a picture of what the current state of research has to offer in terms of technology and investigation methods. After an analysis of the features offered by the tools, a summary is drawn and potential room for improvement in the field is identified.

1 Introduction

In view of the remarkable gains in quality achieved in Machine Translation (MT) in the past years, post-editing machine output is now growing to become an established translation modality in its own right and, as a result of this, an increasing number of researchers are starting to investigate the process of PE.

Research in the field tends to be particularly focused on the effort invested in the activity. However, investigating effort in PE, or in any other task, is a very challenging undertaking. Especially if temporal, technical, and cognitive effort (Krings, 2001) are taken into account, the use of tools that are able to log time, keyboarding, as well as potential indicators of cognitive effort (e.g. gaze data) becomes paramount in achieving research goals.

In order to cast light on the type of data obtained in PE investigations with the use of research tools currently available, four pieces of software are reviewed in the present paper: Tobii Studio (v.3.1.3), Translog-II (v.0.1.0.189), TransCenter (v. 0.5), and PET (v. 2.0). These tools are chosen for analysis due to their prominence in previous research and their possibility of being exploited specifically for PE. All four tools are described in view of key- and time-logging features, and data visualisation aids, while eye-tracking features are only considered in view of Tobii Studio and Translog-II, since TransCenter and PET do not offer a built-in integration with eye trackers.

In describing these tools based on empirical data, the aim of this paper is to provide an overview of the current state-of-affairs in PE research technology and point to potential aspects that can be further improved in the field. Tools such as the ones used in Green et al. (2013) and Plitt and Masselot (2010), as well as the productivity-testing tool available in the context of the TAUS Dynamic Quality Framework[1] are not reviewed here, since, to the best knowledge of the author, they are not available to the general public. With regard to other tools that can be used for PE research, the CASMACAT (Ortiz-Martínez et al., 2012) and MateCat (Cattelan, 2012) workbenches were not included in the present review. Even though prototypes and beta releases of the tools are available, at the time of writing, the tools' development projects are yet to be finalised. Appraise (Federmann, 2012) and iOmegaT (Moran and Beregovaya, 2012), which are mainly focused on MT evaluation and PE productivity measurement, respectively, are two other tools that have not been reviewed. Even though these tools can be used for PE research, due to space and time limitations they could not be included in the analysis.

In the remainder of this paper, the criteria for analysis of the tools, the tasks conducted to investigate their usability, and a brief description

[1] https://tauslabs.com/dynamic-quality

Sharon O'Brien, Michel Simard and Lucia Specia (eds.)
Proceedings of MT Summit XIV Workshop on Post-editing Technology and Practice, Nice, September 2, 2013, p. 93–101.

of each tool are provided in Section 2. The tools are analysed in Section 3, and, in Section 4, suggestions are provided in terms of potential adaptations and areas where there is possible room for improvement in regard to technology that can be applied to PE research.

2 Context and Criteria for Analysis

In terms of the functionalities comprised in the tools that can be useful for research in PE, Tobii Studio and Translog-II are analysed based on the following eye-tracking-specific aspects:

- Amount of information comprised in gaze data logs and ease in computing measures at a sentence and/or sub-sentence level;
- Possibility of measuring gaze data quality.

As to features that do not necessarily involve eye tracking, all four tools are analysed based on the following criteria:

- Amount of information in time and key logs and ease in computing measures at a sentence level;
- Data visualisation aids;
- Customisation possibilities within the tools' environment.

The choice of these specific criteria is motivated by potentially challenging methodological aspects observed in previous research, such as computing per-segment PE time based on timestamps in the task video (O'Brien, 2011), and computing gaze data pertaining to ST and TT windows based on screen pixel positions (Hvelplund, 2011). Gathering UAD at the sentence level seems to be, overall, a common and yet challenging research need, frequently incurring in task designs where sample materials are exposed to subjects sentence by sentence, with no access to the whole text being granted – which is the case in Green et al. (2013) and Doherty et al. (2010), for example. Nevertheless, the criteria chosen in this review are by no means exhaustive, and the question of what exact set of features make for a good research tool in PE cannot be entirely solved in this paper.

The studies conducted to test the tools consisted of PE tasks with source text (ST) in Spanish and target text (TT) (MT output) in English. Spanish news texts of approximately 130 words each were translated into English with Google Translate [2], and two professional translators post-edited the MT outputs. The eye-tracking equipment used with Tobii Studio and Translog-II is a Tobii X120 remote eye tracker.

2.1 Tobii Studio

Tobii Studio is the Windows-oriented eye-tracking software that accompanies Tobii eye trackers. Since the program does not have a built-in text editor, the screen-videoing mode needs to be used for PE tasks. When running in this mode, the program records everything that happens on the computer screen in the format of an .avi video, superimposing individuals' eye movements onto the recording. Data can be manipulated within the tool or exported in .tsv or .xlsx formats. Microsoft Word was the text editor used in combination with Tobii Studio. While this is arguably not the best text editor for PE research, this analysis only concerns features that apply specifically to Tobii Studio. In that way, Microsoft Word editing features and user interface (UI), as well as their usability for PE research, are beyond the scope of this paper.

2.2 Translog-II

Translog-II (Carl, 2012a) is a Windows-oriented software package designed specifically for translation process research (TPR). The package contains two tools: the *Supervisor*, and the *User*. Projects are set in the Supervisor, where any data produced can be visualised and manipulated. The User serves as the editing interface where participants carry out the task. Other than gaze data, the tool can also record keyboard and mouse events, as well as audio. In addition to the analysis possibilities presented within the environment of the tool, Translog-II data log files, which are saved in .xml format, can be further processed by a series of scripts included in the TPR database of the Centre for Research and Innovation in Translation and Technology (CRITT TPR-DB) (Carl, 2012b). Since these scripts are designed to process data in the format obtained with Translog-II, they are also taken into account in the present analysis, which is based on Version 1.2 of the scripts.

2.3 TransCenter

TransCenter (Denkowski and Lavie, 2012) is an open-source, web-based tool that allows different

[2] http://translate.google.com/

participants to carry out PE tasks remotely via a server. The tool logs time and keyboard/mouse activity at a sentence level. Subjective assessments of translation quality, difficulty, and usability can also be gathered through quality rating scales that are automatically included in the tool's UI depending on the task chosen – if bilingual or monolingual PE, for example. Aggregate UAD for all participants, as well as data for each participant individually can then be accessed via report files generated by the tool in both .csv and .html formats. Since the tool is web-based, TransCenter can be accessed on any platform.

2.4 PET

Out of the four tools considered, PET (Post-Editing Tool) (Aziz et al., 2012) is the only one designed specifically for PE. Similarly to TransCenter, PET is open-source and platform-independent. In addition to recording time and effort indicators at a segment level, PET also allows users to perform assessment tasks based on configurable rating scales and criteria. UAD generated with PET is saved in .xml format.

3 Analysis

In the following section (3.1), Tobii Studio and Translog-II are analysed in view of eye-tracking-specific features. Since TransCenter and PET do not log gaze data, these tools are not analysed in this section. In sections 3.2 and 3.3, all tools are taken into account.

3.1 Gaze data

3.1.1 Amount and type of information in gaze data logs

In Tobi Studio's data log file, gaze events are classified as 'fixation', 'saccade' or 'unclassified', and each event is accompanied by information such as the positions of both right and left eyes on screen, left and right pupil sizes, distance of both eyes from the screen, as well as the gaze event's duration in milliseconds. Clusters of gaze events that are identified as a single fixation or saccade receive a respective index number. Gaze events are grouped into fixations based on fixation-filter settings that are configured by the user. An extract of the data log file generated with Tobii Studio is presented in Table 1.

Recording Timestamp	FixationIndex	SaccadeIndex	GazeEventType	GazeEventDuration
430	5		Fixation	158
439		5	Saccade	42

Table 1. Extract of Tobii Studio data log file

As for Translog-II, the .xml results file with UAD contains information such as source and target (MT output) texts, the task (if 'translating' or 'post-editing', e.g.) as well as keyboard, mouse and time logs, cursor positions, and gaze data. In terms of gaze data, the file includes information such as the timestamp associated with each gaze event, positions of right and left eye on screen, as well as pupil size.

With respect to differences between Tobii Studio and Translog-II in terms of the type of gaze data generated, the latter – being a tool specifically designed for TPR – automatically records information pertaining to the particular window (ST or TT) a given gaze event is related to. In Translog-II, gaze events can be filtered into fixations based on the CRITT TPR-DB scripts. After aligning ST and TT with the jdtag[3] tool, these scripts can be used to produce, among other things, a series of unit tables with process and product data as well as files that can be used in external tools for part-of-speech (POS) tagging and syntactic parsing. Below is an example of a fixation data (FD) table generated with the CRITT TPR-DB scripts.

FIXid	Time	Dur	Win	Cursor	ParalK	Edit	EditID	STid	TTid
29	7987	1066	1	878	0	---	---	159	157
30	10389	142	2	148	142	_	10+	28	29
31	10561	366	2	55	94	of	11+	12	12

Table 2. Translog-II Fixation Data

As shown in Table 2, similarly to Tobii Studio, each eye fixation in Translog-II has an individual ID and is accompanied by duration and timestamp (the columns 'Dur' and 'Time', respectively). In Translog-II, however, thanks to the gaze mapping functionality of the tool, it is also possible to know what word in the text the fixation refers to. Each word in the text is given an ID number, and ST and TT word ID pairs (columns 'STid' and 'TTid') associated with fixations are also displayed in the table.

Still in regard to gaze data, also generated by the CRITT TPR-DB scripts is a table with fixation units (FU). A concept proposed by Carl and Kay (2011), FUs are clusters of fixations that, together, represent one meaningful sequence. FU

[3] http://code.google.com/p/jdtag/

tables have information on the time each unit started, its duration, as well as the amount of time for which reading and typing took place in parallel.

Data exported from both Tobii Studio and Translog-II could arguably be deemed to be in interoperable formats overall, since the former exports data in .tsv format, and the latter in .xml. As regards the replay function, however, the fact that tasks can be replayed based on the .xml file in Translog-II arguably allows for an easier storage and transport of data. In Tobii Studio, by contrast, tasks are replayed from .avi files, which tend to be considerably large and hence potentially difficult to store and transport.

3.1.2 Computing gaze-data measures at a sentence and/or sub-sentence level

As to computing measures for specific moments of the task, Tobii Studio allows the possibility of selecting video passages and marking them as 'scenes'. Statistics referring to specific 'areas of interest' (AOIs) on the screen (ST and TT windows, say) within a scene are then computed. In that way, if specific sentences or phrases can be identified in the text as AOIs, it is possible to draw a polygon around the corresponding area and obtain data pertaining only to the particular area selected.

In Translog-II, by contrast, each fixation is automatically mapped to specific ST and TT words with the use of the CRITT TPR-DB scripts. Even though the quality of the gaze mapping might also depend on the precision of the eye tracker used, this functionality arguably allows for an easier consideration of gaze data at a sentence and/or sub-sentence level. In addition, Translog-II offers the possibility of correcting gaze mapping manually after conducting a task – a functionality not offered by Tobii Studio.

3.1.3 Measuring gaze data quality

In experimental designs where eye-tracking data is used, an important step in the analysis process is to account for data quality. In this respect, Tobii Studio has a built-in measure that assesses the confidence that a given gaze event is in fact valid, generating values that can range from 0 (high confidence) to 4 (no eye found).

Measures of data quality can also be computed based on information in Tobii Studio's data log file. In Hvelplund (2011:103-107) e.g., where a previous version of Tobii's eye tracking software was used, mean fixation duration, gaze-sample-to-fixation percentage, and a ratio between gaze time on screen and total production time have been used as indicators of gaze data quality. In Translog-II, a ratio of gaze events happening in windows 1 and 2, and gaze events that did not happen in any window, i.e. events that have 0 as a window value, constitutes another potentially interesting strategy to measure data quality informally suggested to the author by Translog-II developers.

Overall, with respect to gaze data, while Tobii Studio and Translog-II generate raw output files with similar information, the scripts in the CRITT TPR-DB database allow for a number of further automatic data analysis stages which, in Tobii Studio, would arguably involve lengthy processing steps.

3.2 Key and time logs

3.2.1 Amount and type of information in key and time logs

In addition to gaze data, Tobii Studio also logs keyboarding and mouse clicks, which can be found together with gaze data in the same log file. All these events are associated with their respective timestamp based on the task video.

With respect to Translog-II, CRITT TPR-DB unit tables include a keystroke-data (KD) table, as well as production-unit (PU) and alignment-unit (AU) tables. The KD table includes information on the number and type of editing operations performed (insertions, deletions) and the words in the ST and post-edited text associated with them. Similarly to the concept of FU, PU are clusters of editing operations that can be regarded as a single unit. AU tables, in turn, contain process and product data pertaining to aligned source and post-edited units, i.e. the edits performed and the aligned result of these edits.

With respect to TransCenter, measures such as edit, keypress and mouseclick counts are recorded per sentence. The tool also records editing time, and how each sentence is scored by participants based on 1-5 scales.

As regards PET, the tool distinguishes between white-space, non-white-space and control keyboard events, classifying each event according to a fine-grained list of categories, including e.g. 'navigation-keys' and 'paste-keys'. In addition, it offers a few functionalities that are different to the ones found in other tools, such as au-

tomatically labelling clusters of insertions and deletions as 'substitutions' and 'shifts', and computing Human Translation Edit Rate (HTER) (Snover et al., 2006) as a built-in effort indicator. PET also logs sentence/segment-specific measures of editing time.

3.2.2 Computing key and time measures at a sentence level

In terms of time-logging, Tobii Studio simply offers the timestamp associated with each event recorded by the tool. One way of computing measures of time at a sentence level with Tobii Studio is by considering the timestamps in the task video associated with the moments when participants began and finished editing each sentence. In this respect, if the task is not carried out on a sentence-by-sentence nature where each segment/sentence needs to be confirmed before moving on to the next, computing sentence-specific time measures in Tobii Studio constitutes an arguably unreliable approach, since it would be hard to collect such measures without distinct time delimitations between sentences.

With respect to Translog-II data, due to the tool's gaze mapping functionality, information on time can be obtained for each FU or PU, for example. In addition, when setting up an experiment in Translog-II, the ST can be divided into translation units that are displayed separately according to settings established by the researcher, such as a time limit for which the segments will be displayed. The time spent on each unit can then be observed in the data log file that is generated after a task is completed. However, in the context of this evaluation, this functionality did not seem possible to be used for PE, since only the ST seems to be breakable into units, and not both ST and TT (MT output). In this respect, PET and TransCenter seem to be the only tools analysed that offer an automatically computed measure of time per ST-TT segment, which can be useful in PE task designs where sentences are established as units for analysis.

In sum, while time measures at a sentence level need to be computed based on timestamps in Tobii Studio, in Translog-II these measures can be computed for ST-based units, or for sentence and sub-sentence units based on fixations and/or keyboard events. PET and TransCenter, in turn, offer automatically computed key and time measures per ST-TT segment.

3.3 Data visualization aids

In addition to quantitative data that can be exported from Tobii Studio for each task, the software has a number of different graphic representations of data that can be explored. Tables and charts can be generated and gaze events can be viewed in the form of gaze plots, where eye fixations can be observed on a still screen capture extracted for a given timespan in the task video.

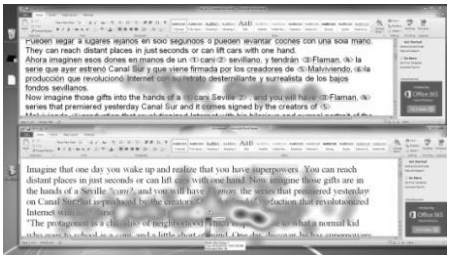


Figure 1. Tobii Studio Heat Map

Another visualisation option offered by Tobii Studio are heat maps (Fig. 1), where a colour representation – ranging from green (cool) to red (hot) – indicates the areas of the screen that received more gaze events.

One of the most prominent visualisation options in Translog-II is what is referred to as the 'linear view' (Fig. 2), where the editing process can be observed linearly with different editing events (including eye fixations), represented by different symbols and colours.

{2:Margare}{2:et That}{2:atcher,}{1:vera de}{1:et That}{1:Margare}{1:primera}{1:
mujer }{1:e alcan}{1:ó el ca}{1: cargo }{1:e prime}{1: minist}{1:
primer}{1:ministr}{1:Reino U}{1:rimer m}{2:tcher, }{2:he firs}{2: woman }{2: to
rea}{2: reach }{2:he prim}{2:me mini}{2:ch the }{2: reach }{2:e prime}{1:as.
}{2:r of Un}{1:as.
[▼][▲]}•post{2:f 1979 •}{2: of prior}{2: minist}{1:as.
}{1:as.
}{2: Thatch}{2:Kingdom}{2:and Bri}{2:ritish }{2:nd Brit}{2:olitica}{2:life tr}{1:.
Según}{1:Margare}{1:y que t}{1: Reino }{1:tro de }{1:minist}{1:e
prime}{1:Margare}{1:Margare}{1:Margare}{1:Margare}{1:as.
}{1:as.
[▼][▲]}◄◄in{2:ctim of•}{2:r in t t}{1:as.

Figure 2. Translog-II Linear View

In the linear view extract in Fig. 2, keyboard, mouse and fixation events are displayed. Portions of fixated text are displayed inside brackets, where the number before the colon represents the window where the fixation occurred – 1 (one) refers to the ST, and 2 (two) refers to the TT. Two consecutive triangles pointing downwards and upwards represent a click. With regard to keyboard events, dots represent spaces, triangles pointing backwards represent deletions, and insertions are displayed simply as the letters that were actually typed by the participant.

Another way of viewing data in Translog-II is by replaying the task via the .xml log file. Data

can also be viewed in the format of a pause plot, where keyboard pauses can be observed in a graph. A screen capture of the replay function in Translog-II is presented in Fig. 3, where the focus of gaze data is represented by a circle and its mapping by a rectangle over the respective portion of text being fixated.

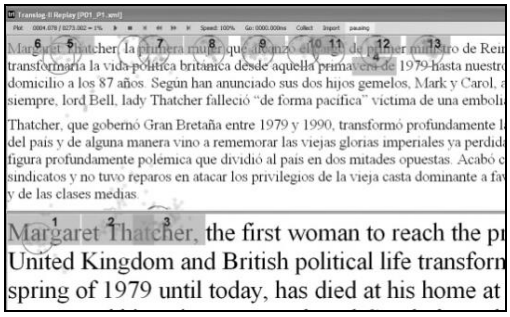


Figure 3. Translog-II Replay Function

Translation progression graphs (TPGs) (see Fig. 5) present another possibility of visualising Translog-II data. A feature exclusive of Translog-II, these graphs can be generated with the statistical package R[4] based on the tables created with the CRITT TPR-DB scripts.

TPGs can be very informative in denoting combined reading and production patterns. Perhaps to make such graphs more useful in the context of PE, adding reference to the post-edited text (and not only the ST) would be desirable.

With respect to TransCenter, the tool enables a sequential edit-by-edit visualisation of the PE process through 'edit trace reports' (Fig.4). The tool also displays aligned ST, MT output and post-edited sentences together with sentence-specific UAD.

Time	Sentence 4 Edits
Initial	How does Mary Shelley finish off?
1370010899990	How does Mary Shelley finish off?
1370010899990	How w Mary Shelley finish off?
1370010900057	How wo Mary Shelley finish off?
1370010900108	How wou Mary Shelley finish off?
1370010900252	How woul Mary Shelley finish off?
1370010900348	How would Mary Shelley finish off?

Figure 4. TransCenter Edit Trace Report

In regard to PET, no pre-set data visualisation options seem to be available within the environment of the tool. In this respect, while TransCenter has interesting visualisation possibilities not offered by PET, the latter seems to provide more detailed keyboard data, which can always be explored by the researcher in external data-analysis tools.

[4] http://www.r-project.org/

Overall, in terms of visualisation possibilities, heat maps figure as a distinctive feature of Tobii Studio, while TPGs constitute a feature that can be especially useful for PE research and which is offered exclusively by the CRITT TPR-DB scripts. A linear view of the editing process is offered by Translog-II, with a similar and less detailed alternative being offered by TransCenter in the form of edit trace reports.

3.4 Customisation Possibilities

In this section, customisation options presented within the environment of the tools are analysed. While PET and TransCenter are both open-source tools, the analysis presented here focuses on settings that can be customized without resource to the tools' source code.

Since, in the context of PE tasks, Tobii Studio needs to be used with an external text editor, the customising possibilities presented by the tool itself are limited to fixation-filter settings and data visualisation options.

In addition to data visualisation options, such as the colour representation and choice of events to be included in the linear view, Translog-II presents a few task-related customisation possibilities, such as choosing reading, translating or writing as linguistic tasks, and having the window panes displayed accordingly. In the replay mode in Translog-II, it is also possible to choose the FixMap option, where gaze mapping can be manually corrected. With respect to the data log file generated with Translog-II and how it can be processed, a number of possibilities are available to the researcher by manipulating the CRITT TPR-DB scripts, including the configuration of fixation-filter settings, which can be recomputed with the command 'remap'.

Being a tool designed specifically for PE, PET presents a number of potentially useful options that can be explored in the specific context of PE research, such as displaying buttons that allow participants to either accept the MT output as is or discard it altogether – actions that can be tracked later in the results log file generated by the tool. PET also has a drag-and-drop functionality that allows text to be moved both within an active unit, as well as from any segment in the text into the active TT unit being edited.

In comparison with PET, TransCenter seems to offer fewer customisation possibilities that can be configured with no recourse to the tool's source code. No instructions were found on how

to change rating scales or the way panes are displayed in the editing interface, for example. On the other hand, TransCenter is the only tool out of the ones analysed that can be accessed via a server. While PET can also be used remotely by participants, being able to access TransCenter with a username and password on a web browser arguably facilitates the data-collection process, which can be controlled remotely by the project's administrator.

PET also allows access to dictionaries and other reference material within the environment of the tool. While this functionality could also be observed for Translog-II in the tool's documentation, this feature did not seem to be included in the version of Translog-II analysed, nor in a subsequent version (v.0.1.0.191) released after the experiments reported in this paper had been conducted. This renders PET the only tool reviewed to have a functional integration with reference materials.

4 Conclusion and Further Issues

A summary of the functionalities observed for each of the tools described can be observed in Table 4.

As can be seen from the descriptions provided, both Tobii Studio and Translog-II allow an analysis of gaze and keyboard/mouse data both quantitatively, with the generation of tables and statistics, and qualitatively, with features such as the replay function, the linear view and TPGs. PET figures as a powerful option mainly for quantitative investigations specifically on PE, presenting pre-set configurable functionalities that are particularly useful for gathering human assessments as well as measuring temporal and technical effort in PE, which, as with TransCenter, can be considered at a sentence/segment level. With regard to TransCenter, one of the main differentials of the tool seems to be the fact that it is web-based, which allows for an arguably easier running of research tasks.

Tobii Studio constitutes an option that can be adopted when research experiments need to be more ecologically valid and not necessarily strictly controlled, since any commercial CAT tool, such as Trados[5] or memoQ[6], can be used in combination with Tobii Studio. In this respect, combining the use of PET or TransCenter with Tobii Studio also figures as a potentially interesting possibility in which all the PE-specific UI functionalities of PET and TransCenter can be exploited in eye tracking studies.

As regards file formats, all four tools seem to meet good levels of interoperability, with data being saved in formats such as .csv, .tsv, and .xml.

Features	Tobii Studio	Translog-II	TransCenter	PET
Pupil size recording	x	x	n/a	n/a
Built-in gaze data validity measure	*x*		n/a	n/a
TPGs		*x*	n/a	n/a
Heat maps	*x*		n/a	n/a
Gaze plots	x	x	n/a	n/a
Quant. info on saccades	*x*		n/a	n/a
Automatic gaze mapping		*x*	n/a	n/a
Manual gaze mapping		*x*	n/a	n/a
Pause plots		*x*		
Linear view		x	x	
Replay function	x	x		
Integr. w/ audio recorder**	x	x		
Integr. w/ reference materials		x*		*x*
Time recording per segment		x	x	x
Human assessments			x	x
Inference on shifts and subs. (keyb.)				*x*
Drag and drop				*x*
Division of text into PE units			x	x
Locking/hiding PE segments				*x*
Server			*x*	
Platform-independent			x	x
Open-source			x	x

*Not observed in v. 0.1.0.189
**Not used in the context of this paper

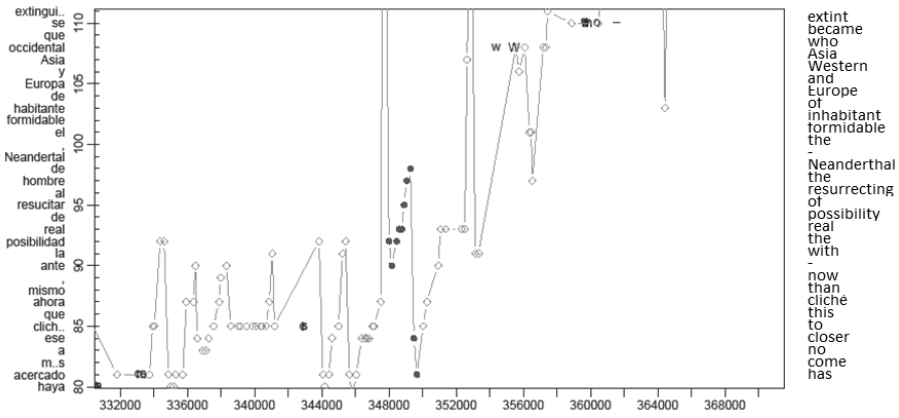
Table 4. Summary of Features

In terms of qualitative analyses, it seems that Translog-II is able to provide a larger number of possibilities to be exploited. In the context of PE, TPGs would arguably be more informative if the post-edited text is also displayed. For demonstration purposes, an adapted version of such graphs is presented in Fig. 5 together with retrospective verbalisations. In these graphs, the y-axis shows ST words in sequence, dark circles represent fixations in the ST, lozenges represent fixations in the TT, black characters represent insertions and red ones represent deletions.

When accompanied by spoken data (in this case, retrospective think-aloud protocols) and the post-edited text, TPGs potentially allow for a powerful and in-depth analysis of the PE process. In this example, it is possible to observe, for instance, that in the time interval shown, the participant had few fixations on the ST relating to the text passage displayed, as signalled by the small

[5] http://www.trados.com/en/
[6] http://kilgray.com/products/memoq

number of dark circles within the range of the graph. It is also possible to observe that the process of editing this passage was far from linear, which is demonstrated by the saccades in the map, where the participant seems to be reading backwards and forwards in an overlapping fash-



Not closer to the cliché, I'm not quite sure, *no closer to the cliché*. I was trying to make sense out of it, that's really why I've added *come. This cliché*, I'm making explicit, perhaps not necessary, but the fact that the Hollywood thing is a cliché. Just capitalising *Western Asia*, I think it's better.

Figure 5. Translog-II translation progression graph with TT and spoken data

ion.

By referring to retrospective spoken data pertaining to the same text passage covered in the graph, it is possible not only to provide a clearer indication of the changes taking place – since deletions and insertions frequently overlap in the graph, hindering full comprehension – but also show the possible mechanisms behind the edits performed.

In terms of other features that could be implemented in tools that can be used for PE research, computing the amount of mouse hovering events and mapping them to their corresponding words in the text figures as a potentially interesting function to be explored. This approach has been used for PE by Green et al. (2013), who mention previous studies where mouse hovering has been shown to correlate with eye-tracking data. In view of the constraints imposed by eye tracking due to the need for specialised equipment and appropriate conditions, it would perhaps be interesting to see automatically computed measures of mouse hovering in freely available research tools that can be used for PE. It is noteworthy, however, that studies looking at the correlation of gaze data with mouse hovering specifically for PE are apparently lacking, which, despite its potential utility, renders debatable the reliability of this measure.

As a relatively new activity, research methods currently available for PE seem to heavily draw on more established areas such as reading and traditional translation. In view of this, it seems that in-depth studies into the operational underpinnings of PE would lead to better strategies of data collection that reflect the PE activity more directly. The amount of crossing between ST and MT output is an example of a potential measure of effort in PE that is arguably underexplored. In addition, it seems that only recently there have been initiatives at developing data-collection tools that are able to mimic more advanced CAT functionalities offered in commercial CAT software, which is an aspect that the CASMACAT and MateCat projects aim to attend to. In this respect, the controlled lab conditions enabled by research tools such as the ones reviewed in this paper might hinder more valid investigations into effort, since, when using these tools, participants are not able to count on functionalities that they would normally be able to use in real-world contexts, such as interactive editing features and on-the-fly quality assurance checkers, for example.

As future work, it would be interesting to expand this review by including the analysis of other tools. Data obtained with other studies could also be considered in order to check to see if research needs are met across a wider and more diverse context.

Acknowledgments

This evaluation is part of a project supported by the School of Modern Languages at Newcastle University. Particular gratitude is extended to Dr Francis Jones, Dr Michael Jin, and Dr Ya-Yun Chen for their collaboration.

References

Aziz, W., Sousa, S. C. M., and Specia, L. 2012. PET: a tool for post-editing and assessing machine translation. *The Eighth International Conference on Language Resources and Evaluation, LREC'12*, Istanbul, Turkey. May 2012, 3982-3987.

Carl, M. 2012a. Translog-II: a program for recording user activity data for empirical reading and writing research. *The Eight International Conference on Language Resources and Evaluation, European Language Resources Association*, Istanbul, Turkey. May 2012, 4108-4112.

Carl, M. 2012b. The CRITT TPR-DB 1.0: A Database for Empirical Human Translation Process Research. *AMTA 2012 Workshop on Post-Editing Technology and Practice (WPTP 2012)*. ed. / Sharon O'Brien; Michel Simard; Lucia Specia. Stroudsburg, PA : Association for Machine Translation in the Americas (AMTA), 2012. 9-18.

Carl, M. and Kay, M. 2011. Gazing and Typing Activities during Translation: A Comparative Study of Translation Units of Professional and Student Translators, *Meta: Journal des traducteursMeta:/Translators' Journal*, 564, 952-975.

Cattelan, A. 2012. MateCat. *D4.1 First Version of MateCat Tool*. Available at http://www.matecat.com/wp-content/uploads/2013/01/MateCat-D4.1-V1.1_final.pdf [Accessed 16 July 2013].

Denkowski, M. and Lavie, A. 2012. TransCenter: Web-Based Translation Research Suite, *AMTA 2012 Workshop on Post-Editing Technology and Practice Demo Session*, 2012.

Doherty, S., O'Brien, S., and Carl, M. 2010. Eye tracking as an MT evaluation technique. *Machine translation*, *24*(1), 1-13.

Federmann, C. 2012. Appraise: An Open-Source Toolkit for Manual Phrase-Based Evaluation of Translations. *The Prague Bulletin of Mathematical Linguistics (PBML)* 98, 25-35.

Green, S., Heer, J. and Manning, C.D. 2013. The Efficacy of Human Post-Editing for Language Translation. *ACM Human Factors in Computing Systems (CHI)*, 2013.

Hvelplund, K.T. 2011. *Allocation of Cognitive Resources in Translation: an Eye-Tracking and Key-Logging Study*. Ph.d.-afhandling. Copenhagen Business School Copenhagen

Krings, H. P. 2001. *Repairing Texts: Empirical Investigations of Machine Translation Post-Editing Processes.* Vol. 5. Kent, Ohio: Kent State University Press.

Moran, J. and Beregovaya, O. 2012. iOmegaT – an adapted open--source CAT tool to measure. MT post--edi ng produc vity in enterprise deployments. Demo Poster in: *AMTA 2012*. Available at http://m25s17.vlinux.de/098709809/AMTA2012_DemoPoster.pdf [Accessed 16 July 2013].

O'Brien, S. 2011. Towards predicting post-editing productivity. *Machine translation*, *25*(3), 197-215.

Plitt, M. and Masselot, F. 2010. A productivity test of statistical machine translation post-editing in a typical localization context. *Prague Bull Math Linguist*, 93, 7-16.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L. and Makhoul, J. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. *The 7th Conference of the Association for Machine Translation in the Americas*, 223-231.

Ortiz-Martínez, D., Sanchís, G., Casacuberta, F., Alabau, V., Vidal, E., Benedí, J. M., González-Rubio, J., Sanchís, A. and González, J. 2012. The CASMACAT Project: The Next Generation Translator's Workbench. *The 7th Jornadas en Tecnologıa del Habla and the 3rd Iberian SLTech Workshop (IberSPEECH)*, 326-334.